Rexercise_TOEFLdata

October 13, 2023

First we are going to load the library readr, and then we are going to load the data we downloaded from the Kaggle website. Remember, you can also do this in a point and click fashion in R studio (i.e., click on import new dataset in the top right pane, and select your file).

[25]:
```r
library(readr)
```

Next we are going to import our data using the read.csv function.

[26]:
```r
admissions_data <- read_csv('Admission_Predict_Ver1.1.csv', show_col_types =␣
  ↪FALSE)
```

NOTE: make sure to write the filepath correctly. Here is a hint: if you open your terminal window on a mac, and you drag and drop the file into your terminal, then the terminal will display the filepath. You can then copy and paste that file path between the brackets in the read_csv function above.

Now we are going to look at the first six rows of the data, just to get a sense of how it looks, using the head() function.

[9]:
```r
head(admissions_data)
```

A tibble: 6 × 9

Serial No. <dbl>	GRE Score <dbl>	TOEFL Score <dbl>	University Rating <dbl>	SOP <dbl>	LOR <dbl>	CGPA <dbl>	Resear <dbl>
1	337	118	4	4.5	4.5	9.65	1
2	324	107	4	4.0	4.5	8.87	1
3	316	104	3	3.0	3.5	8.00	1
4	322	110	3	3.5	2.5	8.67	1
5	314	103	2	2.0	3.0	8.21	0
6	330	115	5	4.5	3.0	9.34	1

Now we are going to print the column names from the data, to make sure we are spelling them correctly. This was the biggest source of problems in class yesterday! For this, we will use the colnames() function.

[14]:
```r
colnames(admissions_data)
```

1. 'Serial No.' 2. 'GRE Score' 3. 'TOEFL Score' 4. 'University Rating' 5. 'SOP' 6. 'LOR' 7. 'CGPA' 8. 'Research' 9. 'Chance of Admit'

Notice that the column names have spaces. This is annoying, because computer languages are not great at processing spaces, so we have to be extra careful to write the column name exactly as it is printed above. So, for example, to write the Chance of Admit variable, we have to enclose it in single back quotations.

LEARNING NOTE: I myself could not remember how to write variable names with spaces in R, so this is something I googled. And I found the answer very quickly in Stack Overflow at this link: https://stackoverflow.com/questions/4551424/how-to-refer-to-a-variable-name-with-spaces

Now that we know how to refer to the variabler correctly, we can write our linear model.

[31]:
```
model <- lm(`Chance of Admit` ~ `TOEFL Score`, data=admissions_data)
```

Next we will print a summary of the model, and also a summary of just the parameter coefficients. From this, you can identify the parameter coefficients, their standard error, and their p-values.

[24]:
```
summary(model)

summary(model)$coefficients
```

```
Call:
lm(formula = `Chance of Admit` ~ `TOEFL Score`, data = admissions_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.31337 -0.04990  0.01310  0.05633  0.20725 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -1.2489882  0.0681317  -18.33   <2e-16 ***
`TOEFL Score`  0.0183850  0.0006346   28.97   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08621 on 498 degrees of freedom
Multiple R-squared:  0.6276,	Adjusted R-squared:  0.6269 
F-statistic: 839.4 on 1 and 498 DF,  p-value: < 2.2e-16
```

A matrix: 2 × 4 of type dbl

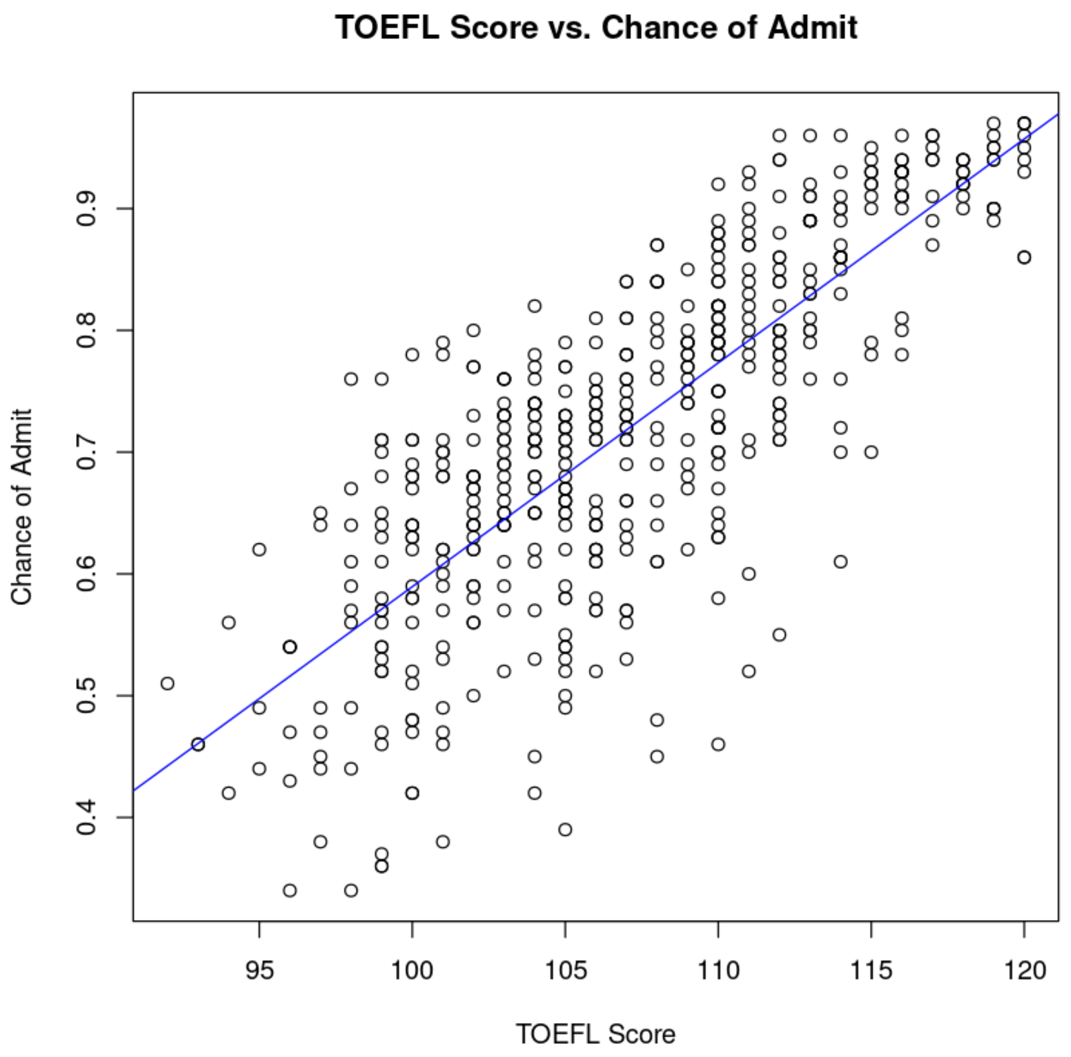
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.24898824	0.0681316826	-18.33197	9.572693e-58
‘TOEFL Score‘	0.01838503	0.0006345856	28.97171	6.729927e-109

Now we can plot our data, as well as the regression line. In this exercise, we did this using the base R plotting function, which is very simple, instead of ggplot2, which will produce the same result. In general, ggplot2 is more flexible, and more fancy, whereas base R plotting is very basic.

[23]:
```
plot(admissions_data$`TOEFL Score`, admissions_data$`Chance of Admit`,
     xlab = "TOEFL Score",
     ylab = "Chance of Admit",
     main = "TOEFL Score vs. Chance of Admit"
```

```r
    )

abline(model, col = "blue")
```


TOEFL Score vs. Chance of Admit
Chance of Admit
TOEFL Score